W586–W589 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh440

COLORADO3D, a web server for the visual analysis of protein structures

Joanna M. Sasin* and Janusz M. Bujnicki

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, Warsaw, Poland

Received February 15, 2004; Revised and Accepted April 16, 2004

ABSTRACT

COLORADO3D is a World Wide Web server for the visual presentation of three-dimensional (3D) protein structures. COLORADO3D indicates the presence of potential errors (detected by ANOLEA, PROSAII, PROVE or VERIFY3D), identifies buried residues and depicts sequence conservations. As input, the server takes a file of Protein Data Bank (PDB) coordinates and, optionally, a multiple sequence alignment. As output, the server returns a PDB-formatted file, replacing the B-factor column with values of the chosen parameter (structure quality, residue burial or conservation). Thus, the coordinates of the analyzed protein 'colored' by COLORADO3D can be conveniently displayed with structure viewers such as RASMOL in order to visualize the 3D clusters of regions with common features, which may not necessarily be adjacent to each other at the amino acid sequence level. In particular, COLORADO3D may serve as a tool to judge a structure's quality at various stages of the modeling and refinement (during both experimental structure determination and homology modeling). The GeneSilico group used COLORADO3D in the fifth Critical Assessment of Techniques for Protein Structure Prediction (CASP5) to successfully identify well-folded parts of preliminary homology models and to guide the refinement of misthreaded protein sequences. COLORADO3D is freely available for academic use at http://asia.genesilico.pl/colorado3d/.

INTRODUCTION

Knowledge of the three-dimensional (3D) structure of a protein is essential to understand how a protein performs its function. Analysis of individual structures can provide explanations for specific biochemical functions and mechanisms at the molecular level, while comparison of multiple structures can give insight into general phenomena such as protein evolution, folding and stability. Protein structures can be determined at high resolution by either experimental methods, such as X-ray crystallography and nuclear magnetic resonance (NMR), or computational analysis (i.e. using bioinformatics tools).

Protein structures derived from either experimental structure solutions or computational predictions are merely models that aim to give as good an explanation as possible for the collected data (1). The quality, quantity and care of the data with which the model was built determine whether the available model is an accurate and meaningful representation of the protein molecule. It is well known that in X-ray crystallography and NMR, and especially in protein structure prediction (which is typically based on very sparse data), errors can be introduced at various stages of the model building process. There have been a number of serious errors documented in literature [reviews: (2,3)]. While inaccuracy at the level of stereochemistry does not necessarily make the protein model worthless, mistracing and frame shifts (misalignment of amino acids with respect to the true position in the fold) can seriously mislead the functional interpretation.

To avoid erroneous structural models, various methods have been developed for protein structure validation. One such method verifies the model's agreement with the original experimental data (or a subset which has not been used for model building). Another set of methods, the so-called knowledge-based methods, examine the geometry, stereochemistry and other structural properties of the model independently from the original data. This latter set of methods reports the extent to which the parameters of the analyzed structure fit within the range of values observed in previously solved high-resolution structures. For optimum performance, several different methods should be used before the structural model can be regarded as validated and assumed error-free.

Knowledge-based methods for protein structure validation typically require the full-atom Protein Data Bank (PDB)-formatted file of protein structure as input, and present the output as a list of confidence values associated with each amino acid or each atom. Poorly modeled amino acids frequently

*To whom correspondence should be addressed. Tel: +48 22 668 5384; Fax: +48 22 668 5288; Email: asia@genesilico.pl

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

occur in regions close in space, which may be distant at the primary sequence level. The graphical presentations of results reported by all structure validation programs are limited to plotting the confidence values against the position in the polypeptide. This output does not allow for the inspection of sites of potential errors at the level of the 3D structure.

We have developed a simple program available as a World Wide Web server, COLORADO3D, which greatly facilitates the visualization of various features directly at the protein structure level with the aid of commonly used viewers such as RASMOL (http://www.umass.edu/microbio/rasmol/) (4) or SWISSPDBVIEWER (http://www.expasy.org/spdbv/mainpage.htm) (5). A very useful feature of COLORADO3D is its ability to visualize, in color, potential errors in protein structures (derived from either experimental analysis or modeling), regions buried in the protein core and inaccessible to the solvent, and regions of high sequence conservation.

METHODS

The COLORADO3D server is freely available for academic users who sign the license agreement at the URL http://asia.genesilico.pl/colorado3d/. Some of the third-party components (including PROSAII and VERIFY3D) may be unavailable for commercial users, who are nevertheless welcome to contact us to obtain a separate, but limited, license or to obtain a stand-alone version of the program without the third-party components.

As input COLORADO3D takes a single protein structure PDB file, and as output it returns a PDB-formatted file to the user by email. In the returned PDB-formatted file, the original temperature factor (B-factor) value for each amino acid is replaced by a derivative of a score calculated by one of the methods used by COLORADO3D for protein structure analysis. Henceforth, the artificial temperature factor value in the output file will be referred to as the 'T-factor'. All T-factors are linearly scaled to a range between 00.00 and 99.99, which corresponds to the color spectrum from blue to red when visualized with macromolecular structure viewers such as RASMOL, CHIME and SPDBV. By default, residues with high scores (good structure, highly conserved, deeply buried) will be colored in blue (low T-factor) and residues with low scores (poor structure, variable, exposed) will be colored in red (high T-factor). The intermediate values will range from green to yellow to orange.

COLORADO3D allows users to choose from three types of analyses: structure validation, residue depth and sequence conservation.

Structure validation

COLORADO3D carries out protein structure validations using third-party programs: VERIFY3D (6), PROSAII (7), ANOLEA (8) and PROVE (9). VERIFY3D, PROSAII and ANOLEA are based on the inverse folding approach and evaluate the environment of each residue in a model with respect to the expected environment as found in the high resolution X-ray structures. VERIFY3D operates on the 3D–1D profile of a protein structure proposed by Eisenberg and co-workers (6), which includes the statistical preferences for the following criteria: (i) the area of the residue that is buried; (ii) the fraction of side-chain area that is covered by polar atoms (oxygen and nitrogen); (iii) the local secondary structure. For structures determined by X-ray crystallography, the default option in VERIFY3D is to assess the compatibility of each amino acid residue with the local 3D structure by averaging the 3D–1D score in a window of 21 residues. For protein models, however, we found that the best results are obtained when a shorter window range of 5–11 residues is used. Similarity to VERIFY3D, PROSAII relies on the empirical energy potentials derived from the pairwise interactions observed in well-defined protein structures. This program is, however, more stringent than VERIFY3D. Regions with small structural errors (such as imperfect pairing of hydrogen bonds in the neighboring beta-strands or poor geometry not allowing for proper salt bridge formation) are often acceptable to VERIFY3D (colored green to light blue by COLORADO3D), while PROSAII tends to pin these regions down (colored red by COLORADO3D) (our unpublished observations). Scores reported by ANOLEA combine a pairwise distance-dependent non-local energy term with an accessible surface energy term. It is reported that ANOLEA can identify bona fide errors in models which have been validated as essentially error-free by VERIFY3D and PROSAII (8). Last, COLORADO3D colors the protein structure according to the stereochemical criterion implemented in PROVE, namely the regularity (or irregularity) of the atom volume (9).

The user can submit a single PDB file to be 'colored' by T-factors or a file with multiple structures (for instance a 'project' file with superimposed structures exported from SWISSPDBVIEWER). All individual structures will be validated and colored separately. As already mentioned, this analysis allows for the identification of regions of dubious or unusual structure that may not necessarily be adjacent at the amino acid sequence level, but that are adjacent at the 3D structure level. Comparison of different structures, such as alternative models of the same protein, can greatly facilitate the choice of optimal fragments or versions and, hence, can greatly facilitate the refinement of protein structures. During the last Critical Assessment of Techniques for Protein Structure Prediction (CASP5) experiment, COLORADO3D was used to identify the best potential homology model among a large number of alternative structures as well as to construct hybrid models (by removing 'red' regions and merging 'blue' ones). With the aid of COLORADO3D we were able to construct very good models, which placed us among the 'winners' of the Comparative Modeling category (10). An example of the application of COLORADO3D to identify and remove errors in a homology model is shown in Figure 1.

In addition to submitting a PDB file, users can also optionally include a multiple alignment of homologous sequences, starting with the query sequence. In this case, additional structural models will be generated based on the backbone of the query, with amino acid substitutions modeled by SCWRL (11) according to the equivalencies of residues obtained from the multiple sequence alignment. The selected program (VERIFY3D, PROSAII, ANOLEA or PROVE) will then validate both the query and all the models and return only the query structure, 'colored' according to the average score of all homologous residues for every column in the alignment file. This analysis allows for the verification of the multiple sequence alignment with the structural requirements of the

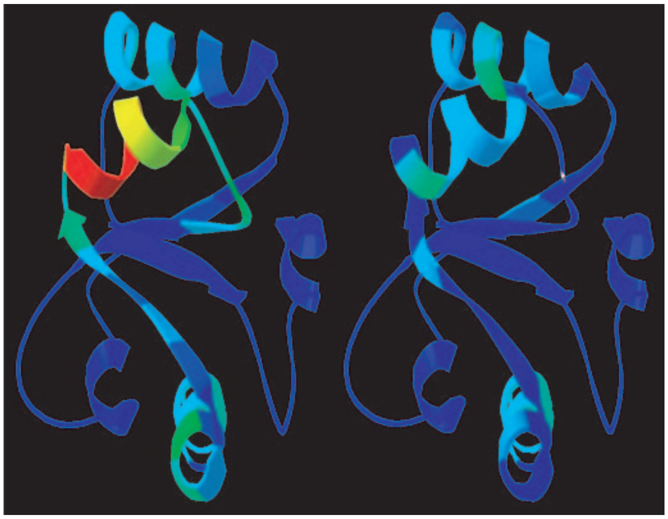

Figure 1. Two alternative homology models of a protein (experimentally solved structure deposited as 1h7m in the Protein Data Bank), colored according to the results of structure evaluation using the VERIFY3D method. Left panel: analysis of the initial model with VERIFY3D via the COLORADO3D server revealed a potentially erroneous region (colored in red), while other parts of the model appear correct (from blue to cyan). Right panel: the model has been corrected (amino acids of one helix were shifted along the polypeptide by one residue, thereby increasing one loop and shortening the other, and changing the interactions of the side-chains with the rest of the protein), leading to the improvement of scores for the suspicious region. The latter model turned out to be indeed correct.

protein fold (and correction of errors in the alignment, if necessary).

Residue depth

The accessible surface area is a parameter widely used in the analysis of protein structure and stability. However, it does not distinguish between atoms immediately below the protein surface and atoms in the core of the protein. COLORADO3D differentiates between such buried residues by utilizing a method developed by Chakravarty (12). The method can be simply described as an estimation of the burial of the hydrophobic surface area. The residue depth parameter is believed to correlate better than accessibility with the effects of mutations on protein stability and on protein–protein interactions. In COLORADO3D, the depth of an atom in a protein corresponds to its distance from the nearest surface water molecule. The depth of a residue is the average of the constituent atom depths.

Sequence conservation

In addition to the visualization of calculated protein structure features, COLORADO3D also allows users to take advantage of structural information to analyze the sequence information in a 3D context. To color the protein structure according to the degree of sequence conservation, the user must submit a single PDB-formatted structure (a template) and a multiple sequence alignment starting with the template sequence. For each column of the alignment, the amino acid percentage identity (i.e. frequency of occurrence of the highest represented amino acid) is taken as a crude measure of the sequence conservation and this percentage is converted into T-factors (linearly scaled between 00.00, for 100% identity, and 99.99, for $\leqslant$ 20%). Using COLORADO3D to color the structure by sequence conservation can greatly facilitate the identification of conserved patches on the protein surface, which may correspond to functionally important sites (such as catalytic centers in enzymes). This analysis is highly complementary to the aforementioned procedure of coloring the structure based on the alignment, but using the average structural score instead of sequence similarity. The two analyses can be used interchangeably to refine the sequence alignment and to thread the sequence along the polypeptide chain. A good example of such a situation is the process of fitting an amino acid sequence into the electron density at the early stages of protein structure determination, when the amino acid identities are still uncertain.

SUMMARY

The COLORADO3D server is a simple, yet useful online tool that allows for the visualization of the results of different protein structure analyses. One big advantage of COLORADO3D over other servers for protein structure 'coloring' is that it provides the user with an output file in PDB format, which can then be reopened and displayed in various orientations and different rendering modes in any molecular structure viewer. During CASP5, we used COLORADO3D to color many alternative structures for each protein according to the result of structure evaluation with VERIFY3D, and in many cases we were able to identify correctly folded parts in imperfect preliminary models. We hope that the present version, enhanced with additional tools, will be even more useful for structural biologists, experimentalists and homology modelers alike.

ACKNOWLEDGEMENTS

We are grateful to the developers of the protein structure analyses included in our meta-server, in particular Professor David Eisenberg, Professor Manfred Sippl, Dr Francisco Melo Ledermann, Dr Joan Pontius and Dr Raghavan Varadarajan. We would also like to thank other members of our laboratory for testing the server. This work was supported by the Polish State Committee for Scientific Research (grant PBZ-KBN-088/P04/2003 to J.M.B.). J.M.B. gratefully acknowledges the support of the EMBO and HHMI Young Investigator Programme and a fellowship from the Foundation for Polish Science.

REFERENCES

1. Laskowski,R.A. (2003) Structural quality assurance. *Methods Biochem. Anal.*, **44**, 273–303.
2. Brändén,C.I. and Jones,T.A. (1990) Between objectivity and subjectivity. *Nature*, **343**, 687–689.
3. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
4. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
5. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
6. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

7. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
8. Melo,F., Devos,D., Depiereux,E. and Feytmans,E. (1997) ANOLEA: a www server to assess protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 187–190.
9. Pontius,J., Richelle,J. and Wodak,S.J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, **264**, 121–136.
10. Kosinski,J., Cymerman,I.A., Feder,M., Kurowski,M.A., Sasin,J.M. and Bujnicki,J.M. (2003) A 'Frankenstein's monster' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, **53** (Suppl.), 369–379.
11. Bower,M.J., Cohen,F.E. and Dunbrack,R.L. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.*, **267**, 1268–1282.
12. Chakravarty,S. and Varadarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure Fold. Des.*, **7**, 723–732.